Significant Properties in the Preservation of Relational Databases

Ricardo André Pereira Freitas[1]
Advisor: José Carlos Ramalho[2]

[1] Universidade Lusíada de Vila Nova de Famalicão,
CLEGI - Centro Lusíada de Investigação e Desenvolvimento em Engenharia e Gestão Industrial
[2] Department of Informatics - University of Minho
Portugal
freitas@fam.ulusiada.pt, jcr@di.uminho.pt

Abstract. Relational Databases are the most frequent type of databases used by organizations worldwide and are the base of several information systems. As in all digital objects, and concerning the digital preservation of them, the significant properties (significant characteristics) must be defined so that adopted strategies are appropriate. In previous work a neutral format (hardware and software independent) — DBML — was adopted to achieve a standard format used in the digital preservation of the relational databases data and structure. Currently, in this PhD project we walk further in the definition of the significant properties by considering the database semantics as an important characteristic that should also be preserved. For the representation of this higher level of abstraction we are going to use an ontology based approach. We will extract the entity-relationship model from the DBML representation and we will represent it as an ontology.

Key words: Digital Preservation, Significant Properties, Significant Characteristics, Relational Databases, Ontology, OAIS, XML, Digital Objects

1 Introduction

In the current paradigm of information society more then one hundred exabytes of data are already used to support our information systems [15]. The evolution of the hardware and software industry causes that progressively more of the intellectual and business information are stored in computer platforms. The main issue lies exactly within these platforms. If in the past there was no need of mediators to understand the analogical artifacts today, in order to understand digital objects, we depend on those mediators (computer platforms). Nothing can guarantee the continuity of access to digital artifacts in their absence [13] and therefore several researchers and research projects aim to face this problem.

Although digital information can be exactly preserved in its original form by only copying (preserving) the bits, the problem appears when we notice the

very fast evolution of those platforms (hardware and software) where the bits can be transformed into something human intelligible [9]. Digital archives and digital libraries are complex structures that without the software and hardware – which they depend on – the human being, or others, will certainly be unable to experience or understand them [8].

Our work addresses this issue of Digital Preservation and focuses on a specific class of digital objects: Relational Databases [9]. Relational databases are a very important piece in the global context of digital information and therefore it is fundamental not to compromise its longevity (life cycle) and also its integrity, liability and authenticity [18]. These kind of archives are especially important to organizations because they can justify their activities and characterize the organization itself. Current studies claim that 90% of the information produced in a daily basis is stored in a relational database.

At this stage of the PhD work we aim to determine/establish the significant properties for the relational databases family of digital objects. First, in the following section, we intend to generally discuss the significant properties for preservation of digital objects and also mention the controversy surrounding the discrepancy of terms used in different ways by different authors [1] (significant properties or significant characteristics). In section 3 the relational databases class of objects must be deeply analyzed; we should be able to completely characterize these type of digital objects so that one may choose what are the issues (the things) important/valid/necessary for preservation. Section 4 establishes the significant properties for relational databases digital preservation. We define a methodology that lead us to the identification of the properties necessary to ensure the preservation of these objects over time. The significant properties are addressed, individually and globally, over different levels of abstraction. At the end we will draw some conclusions, specify the future work to be done and also enumerate some questions that emerge from the research.

2 Digital Objects and Significance for Preservation

The core concept is the fact that digital objects have several associated aspects that we should consider whether or not to preserve. This already divides the scientific community but there is also a discussion surrounding the terms that should be used to address those aspects of the digital objects that should be preserved. Some will defend the terms "significant properties", others use "significant characteristics" and so one [1]. Here we will use mainly the terms "significant properties" when addressing to

> *"The characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects, and their capacity to be accepted as evidence of what they purport to record"*[26].

Angela Dappert and Adam Farquhar state that the "Significance Is in the Eye of the Stakeholder" [1]. However this approach can lead to some confusion because the stakeholder vision of the problem may not always correspond

to the significant properties identified by a community of interest (designated community). The discussion is open!

The perspective of those who intend to develop an action of preservation, over a certain artifact, will restrain/determine the significance of the properties. However, in order to be rigorous and address the problem in its whole essence, the analysis and determination of the significant properties cannot depend on the probable ambiguity of perspectives. Since our study already restrain the family of digital objects addressed for digital preservation — Relational Databases —, there must be a standard that determines the main characteristics/properties to ensure preservation within this class of objects.

2.1 The Digital Object

There can be some distinction between digital objects that already born in a digital context, and those that appear from the process of digitization: analog to digital. In a comprehensive way and encompassing both cases above, we can consider that a digital object is characterized by being represented by a bitstream, i.e., by a sequence of binary digits (zeros and ones) [8].

We can discuss if the physical structure of the object is important, and if so, think about possible strategies for preservation at that level. Nevertheless, there exists a next layer — the logical structure or logical object—, which corresponds to the string of binary digits. They have a certain distribution that will define the format of the object, depending on the software that will interpret it. The interpretation by the software, of the logical object, will provide the appearance of the conceptual object, that the human being is able to understand (interpret) and experiment (Fig. 1).

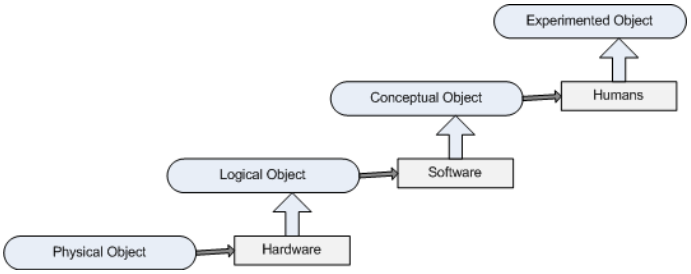

Fig. 1. Digital Object Levels of Abstraction [8]

The strategy of preservation is related to the level of abstraction considered important for the preservation [23]. From a human perspective one can say that what is important to preserve is the conceptual object (the one that the humans are able to interpret). Other strategies defend that what should be preserved is the original bitstream (logical object) or even the original media.

At this stage it is important to beware of a) the relationships established between the levels of abstraction in the digital object and b) that the existence of such chain of relationships is crucial for preservation. If a breach or failure occurs in the chain, the digital object most certainly cease to be intelligible, which may result in the danger of losing the object forever [8].

Some relevant issues stand out: should the environment be preserved? should we question everything? is it feasible? and the answer is that we must be focused; a scientific/specific method must be developed/followed to address these questions.

2.2 Significant Properties

When we have an artifact for preservation — the preservation object or the digital object — in our case relational databases, we could question the effects of cutting/extracting the object from its original context. Can we do this even when we are referring to objects that are platform (hardware/software) dependent? The interaction between the source of the digital object and the platform results on a conceptual object that can be different if the platform changes [26]; The output can be different. The important is that the essential parts purport what the object where made for. So either the source or the platform can be altered if what is essential is obtained maintain the meaning of the digital object over time. The "essential" here is translate into significant properties. In order to be rigorous three things must be defined:

- the artifact for preservation;
- what are all the implications that should be preserved to guarantee the preservation;
- what actions should be taken.

The digital artifact must be exhaustively analyzed in order to be completely characterized. Then a consensus must be establish over what should be preserved. This approach concerns on the analysis about two different levels of abstraction. The higher level of abstraction focuses on the conceptual model and on the semantic of the database. At a more lower level of abstraction we address the structure and data present on the logical model of the database.

3 Relational Databases Properties

A database can be defined as a structured set of information. In computing, a database is supported by a particular program or software, usually called the Database Management System (DBMS), which handles the storage and management of the information. In its essence a database involves the existence of a set of records of data. Normally these records give support to the organization information system; either at an operational (transactions) level or at other levels (decision support - data warehousing systems).

It is fundamental understand and characterize this class of digital objects in order to establish the significant properties that should be preserved. We will try to achieve some consensus over these issues by first characterize theses objects in its whole essence.

3.1 Database Semantics

As we previously mention the information system in several organizations is supported by a database system of some kind. If we intend not only to preserve the data but also the structure of the organization information system we should endorse efforts to also characterize the information system (IS). In other words we must define the conceptual model of the IS.

We are talking about a vision that humans have about a certain IS. The way to specify a conceptual model can be done through an ontology based approach. The ontology of the database should be able to capture the database semantics. This is indeed a significant "property" of the database. If we want to preserve this property, the level of abstraction in terms of significant properties for preservation must be higher.

3.2 Database Data and Structure

The structure of relations and relationships between entities within a database depends on the database model type. Our study focuses on the relational model, widely available and certainly the most used. However there are other logical models for databases: the flat model, the hierarchical model, object-oriented model, among others [27].

Information that indicates the original operating system and the DBMS that used to support the database is important to characterize the environment of the original database. The date of creation of the database and identification of its creator should also be preserved. This information is identified as technical metadata.

The information in a relational database has a particular structure based in relations between tabular data sets usually called tables [5]. Lee Buck [3] and Ronald Bourret [2] on their approaches concerning XML and Databases do not mention any information about the database structure. However, the structure may provide a way of interpreting the data in order to work and extract valid information – knowledge. On one hand we have the data stored in the database and on the other hand its structure. The data contained in the records of the database obviously has to be preserved but through this analysis we conclude that it will be necessary to also preserve the structure of the database [18]. Some structure features considered important for preservation are:

- **TABLES** (Name)
- **COLUMNS** (Name, Type, Size, ...)
- **KEYS** (Primary keys, Foreign keys, ...)

By preserving these elements we are able to preserve all the database structure – relations (tables) and the relationships between them.

There are other features in a database that we should consider whether or not to preserve:

- triggers, functions, stored procedures, forms,
- database users, users privileges,
- jobs, etc.

These elements differ from the previous ones and some of them maybe included in the ontology of the relational database.

4 Significant Properties of Relational Databases

In general the significant properties in a digital object are those that are identified by its community of interest.

Considering the nature of the digital artifacts that we are addressing – relational databases – there is an European strategy encompassed in the "Planets Project" [16] to enable their long term access. The project adopted the SIARD [21] solution, which is based on the migration of database into a normalized format (XML – eXtensible Markup Language [28]). The SIARD was initially developed by the Swiss Federal Archives (SFA).

Another approach, also based on XML, relies on the main concept of "extensibility" – XML allows the creation of other languages [19] (it can be called as a meta language). The DBML [11] (Database Markup Language) was created in order to enable representation of both **DATA** and **STRUCTURE** of the database. The DBML format was adopted by the Portuguese National Archives under the RODA project [20]. The following diagram (Fig. 2) reflects the schema for this language.

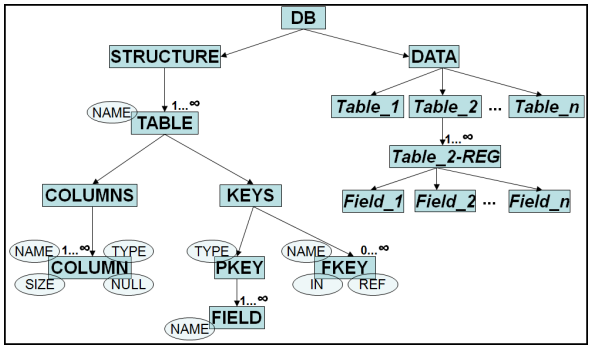

Fig. 2. DBML Schema

Both approaches (SIARD and DBML) adopt the strategy of Migration of the database to XML, why? A neutral format that is hardware and software

(platform) independent is the key to achieve a standard format to use in digital preservation of relational databases. This neutral format should meet all the requirements established by the designated community of interest.

4.1 Previous Works

In the developed work, and considering the preservation of relational databases, we adopted an approach that combines two strategies and uses a third technique: migration and normalization with refreshment [10]. The main strategy in our approach is Migration which is carried in order to transform the original database into the new format – DBML [11]. The normalization is crucial to reduce the preservation spectrum to only one format. A third technique (refreshment) will also be needed. The refreshment consists on ensuring that the archive is using media appropriate to the hardware in use throughout preservation [9].

Until now we have developed a prototype based on a web application with multiple interfaces. These interfaces have the mission to take a certain database and ingest it into the archive. The access to the archive in order to do all the necessary interventions on the system will also be done through those web interfaces [10]. Conceptually, the prototype is based on the Open Archival Information System (OAIS) [4] reference model.

The OAIS model of reference is concerned about a number of issues related to digital preservation: the process of information Ingestion into the system, the information storage as well as its administration and preservation, and finally information access and dissemination [6] [12].

However, the OAIS model does not impose any computer platforms, development language, database management systems (DBMS), interfaces, i.e., does not condition the development of the system at the technological level involved. Instead, the model acts as a guide for those who wish to develop digital archives [4].

The Prototype implementation was a crucial phase of the work. We intend to implement a system capable of ingesting databases, in the form of information packages (DBML + metadata), for preservation. The developed system is based on a Web application and has multiple interfaces that allows not only the ingestion of information, but also its administration, preservation and dissemination. The several Web interfaces can only be accessible through a previous authentication on the system. The administration component manages these requests, and the various privileges with regard to the handling of information in the archive. It is important to refer that so far the work aimed to test the feasibility of relational database digital preservation using this approach. This was indeed possible, i.e., the objective of converting relational databases (different DBMS) into DBML was achieved. We were also able to rebuild the database in a DBMS from the DBML document in order to achieve the database dissemination.

This approach of the problem was used according to the class of digital objects that we addressed – Relational Databases. If the goal was the implementation of a repository for other family of digital objects the strategies may differ [7] [24].

4.2 Policy of Preservation - Current Work

In order to walk further on the main topic of our PhD Project — "Digital Preservation of Relational Databases" —, we intend to also walk further on the determination of the significant properties for this class of digital objects.

After characterizing (section 3) the relational databases digital objects and establish a division between two levels of abstraction, we need to materialize those ideas into packages of information. These packages are to be used as in the OAIS reference model [4].

By focusing our strategy/policy on two levels of abstraction we intend to preserve the two correspondent levels of abstraction present on the chain of relationships of digital objects.

The database **Data** and **Structure**, which we identified as significant properties of the database, correspond to conceptual level of this family of digital object. The migration to DBML covers these properties and ensures that its representation becomes neutral.

At the top of the chain of relationships present in digital objects we have the Experimented Object (interpreted by humans). At this level there is an inherent **Knowledge** associated to the database semantics. We intend to captured the experimented object (knowledge) through an ontology based approach.

Formally,

$$PhisicalObject + Hardware = LogicalObject \tag{1}$$

$$LogicalObject + Software = ConceptualObject \Longleftrightarrow DBML \tag{2}$$

$$ConceptualObject + Humans = ExperimentedObject \Longleftrightarrow Ontology \tag{3}$$

The ontology approach is adopted to formalize the knowledge present at the experimented object level and also a methodology to create an abstract representation of it.

The research work brought us to a point where we seek to preserve the combination of these levels of abstraction. The main strategy in our approach continues to be Migration which is carried in order to transform the original database into the new format – DBML + Ontology.

5 Conclusion and Future Work

Digital preservation is essential to ensure a future access to digital information legacy. From the several different types of digital objects our study focus on the relational databases family.

A prototype was developed to separate the data from its specific database management environment. The prototype follows the OAIS reference model and uses DBML neutral format for the representation of both DATA and STRUCTURE of the database.

At present time, in this PhD project, we address the problem of relational databases digital preservation by pointing at the significant properties of this class of digital objects. A combined strategy is being adopted to integrate as significant properties both conceptual and experimented levels of the digital object. By doing so we intend to provide a neutral (DBML) and abstract (ontology) representation of relational databases.

Some questions emerge during the research for which we seek feedback from the scientific community:

- Concerning databases, what other significant properties should be preserved?
- What other strategies exist to address the problem of conceptual model representation (beyond ontologies)?
- Possible ways to disseminate databases?

In future work we aim to automate the capturing process of the database semantics through an ontology. The integration between OWL Web Ontology Language [14] and Semantic Web Rule Language (SWRL) [22] should provide asserted and inferred knowledge about the database and its information system.

References

1. Angela Dappert and Adam Farquhar, "Significance Is in the Eye of the Stakeholder, " The British Library, Wetherby, West Yorkshire, 2009
2. Ronald Bourret, "XML and Databases, " Copyright 1999-2005 by Ronald Bourret. Last updated September, 2005
3. Lee Buck. "Data models as an XML Schema development method", XML 99, Phyladelphia, Dec. 1999.
4. Consultative Committee for Space Data Systems. "Reference Model for an Open Archival Information System (OAIS) - Blue Book," National Aeronautics and Space Administration, Washington, 2002.
5. Edgar Codd, "A Relational Model of Data for Large Shared Data Banks," in Communications of the ACM, 1970.
6. Michael Day, "The OAIS Reference Model," Digital Curation Centre UKOLN, University of Bath, 2006
7. Claire Eager, "The State of Preservation Metadata Practices in North Carolina Repositories," Chapel Hill, North Carolina, 2003

8. Miguel Ferreira, "Introdução à preservacao digital - Conceitos, estratégias e actuais consensos," Escola de Engenharia da Universidade do Minho, Guimarães, Portugal, 2006.
9. Ricardo Freitas, "Preservação Digital de Bases de Dados Relacionais," Escola de Engenharia, Universidade do Minho, Portugal, 2008
10. R. Freitas, J. Ramalho, "Relational Databases Digital Preservation," Inforum: Simpsio de Informtica, Lisboa, Portugal, 2009, ISBN: 978-972-9348-18-1; [Online]. Available: http://repositorium.sdum.uminho.pt/handle/1822/9740
11. M. Jacinto, G. Librelotto, J. Ramalho, P. Henriques, "Bidirectional Conversion between Documents and Relational Data Bases," 7th International Conference on CSCW in Design, Rio de Janeiro, Brasil, 2002.
12. B. F. Lavoie, "The Open Archival Information System Reference Model: Introductory Guide," Digital Preservation Coalition, Dublin, USA, Technology Watch Report Watch Series Report, 2004.
13. K.-H. Lee, O. Slattery, R. Lu, X. Tang and V. McCrary, "The State of the Art and Practice in Digital Preservation," Journal of Research of the National Institute of Standards and Technology, vol. 107, no. 1, pp. 93-106, 2002.
14. "OWL - Web Ontology Language" [Online]. Available: http://www.w3.org/TR/owl-features/
15. Pat Manson, "Digital Preservation Research: An Evolving Landscape," European Research Consortium for Informatics and Mathematics - NEWS, 2010.
16. "PLANETS - Preservation and Long-term Access through NETworked Services" [Online]. Available: http://www.planets-project.eu/
17. J. Ramalho, M. Ferreira, R. Castro, L. Faria, F. Barbedo, L. Corujo, "XML e Preservação Digital," Dep. Informática, Universidade do Minho e Instituto dos Arquivos Nacionais, Torre do Tombo, 2007
18. J. Ramalho, M. Ferreira, L. Faria, R. Castro "Relational Database Preservation through XML modelling, " Extreme Markup Languages 2007, Montréal, Québec, 2007
19. J. Ramalho, P. Henriques, "XML and XSL - Da Teoria à Prática," FCA - Editora Informática, 2002.
20. "RODA - Repository of Authentic Digital Object" [Online]. Available: http://roda.dgarq.gov.pt/
21. "SIARD - Format Description," Swiss Federal Archives - SFA, 2008.
22. "SWRL: A Semantic Web Rule Language Combining OWL and RuleML" [Online]. Available: http://www.w3.org/Submission/SWRL/
23. K. Thibodeau, "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years," presented at The State of Digital Preservation: An International Perspective, Washington D.C., 2002.
24. D. Waters, "Good Archives Make Good Scholars: Reflections on Recent Steps Toward the Archiving of Digital Information," 2002
25. C. Webb, "Guidelines for the Preservation of Digital Heritage," United Nations Educational Scientific and Cultural Organization - Information Society Division, 2003.
26. A. Wilson, "Significant Properties Report," InSPECT Work Package 2.2, Draft/Version 2 (2007), [Online]. Available: http://www.significantproperties.org.uk/documents/wp22_significant_properties.pdf.
27. Wikipedia contributors, "Database models," in Wikipedia, The Free Encyclopedia, 2008. [Online]. Available: http://en.wikipedia.org/wiki/Database_models/
28. XML, "Extensible Markup Language", in W3C - The World Wide Web Consortium [Online]. Available: http://www.w3.org/XML/